

图计算笔记

面向图计算的加速器

核心问题

- 计算核心低效
 - 指令并行度低
 - 分支预测机制低效
 - 多核拓展性不强
- 存储效率受限
 - 访存并行度低
 - 访存粒度固定
 - 缓存机制低效、

图加速器设计

- 基于FPGA的图加速器（可编程逻辑门阵列）
 - 片上访存优化
 - 片外访存优化
- 基于ASIC的图加速器（专用定制电路）
 - 缓存设计
 - 流水线设计
 - 并行设计
- 基于新型器件的图加速器
 - 混合内存立方体（HMC）
 - Tesseract
 - GraphP
 - GraphPIM
 - 电阻式忆阻器（ReRAM）
 - GraphR

学术界

- Graphicionado（英特尔）
- FPGP（清华）
- ForeGraph（清华）
- TuNao（中科大）

未来挑战

- 易用性
 - 编程模型及支撑库的开发，目前要写硬件层的细节，很难通用，开发效率低。

- 多属性动态图
- 广泛的显示应用
 - 不仅仅局限于社交网络，知识图谱等，还有更多电力国防金融等问题

总结

1. 这篇更关注的是“器”，从硬件底层的方面讲加速和改进，讨论的问题也都是针对于此。
2. 原有的底层设计没有针对于图数据以及图算法做优化。
3. “硬件+模型+算法”一起来解决
4. 这个领域还是需要时间经验积累才能出成果，可视化处于这些偏后的位置。
5. 底层图的渲染遍历事件响应等操作和功能，网上提供的更多的是包装后的接口，应该直接可以兼容图的硬件加速，不兼容的话，这些工具的工程量很大（特别是能够开放出去，具有后面的可维护性）。
6. 后续可以调研下学术界的系统，看下他的抽象层，接口，可视化界面是什么样的，是通用还是要自定义配置。

分布式图计算：从数据划分到任务执行

图计算的特点与挑战

- 很高的访存计算比
- 数据局部性差
- 类型与操作多样
- 规模巨大，结构不规则

图计算的一般模型

- 一般模型（就是普通的计算框架）
- 分布式图计算
 - 整体同步并行计算（BSP）
 - Pregel
 - Giraph
 - 收集-应用-扩散（GAS）
 - PowerGraph
 - 数据流模型（dataflow）
 - GraphX

图数据划分

- 一维划分，点均匀分给机器，每个点存自己的边，负载不均衡
 - Pregel
- 二维划分，边均匀分给机器，划分速度划分效果难两全
 - PowerGraph
- 混合划分，设定阈值，两头分配
 - PowerLra
- 三维划分

未来展望

- 语义更为丰富的编程模块或接口
- 对动态图和时序图的支持
- 环境感知的联合数据划分与计算

总结

1. 这篇讲模型，比硬件的层次高了一点，还是处于略下层，都是为了解决图计算中，负载均衡等方面的问题。
2. 这两篇文章都提到了，编程模型和接口，动态图。可以看到动态图很火热，问题特殊，挑战大，以及现在的许多模型和硬件都是针对于基本图，简单图，在这些方面还没有做到完美，对动态图的支持还不是很好。编程模型上接口方面，去应用图计算的模型框架，需要大量的周边库进行支持才可以更好的拓展到日常。

图计算在阿里巴巴的应用与挑战

挑战

- 分布式，并行图遍历。充分利用遍历路径的并行性加速查询。
- 渐进式计算。递进式的计算两点间的路径，先返回部分“更具代表性”的结果，同时并行探索更多路径。
- 大图可视化。压缩，摘要。

模式匹配

- 异常与反作弊检测
- 实施欺诈拦截
- 挑战
 - 剪枝/并行匹配：针对数据特点制定算法
 - 索引：“热点索引”
 - 增量计算：复用结果
 - 硬件加速

图算法和机器学习

- 构图计算
- 挑战
 - 语言表达能力：对复杂图过程的描述和支持
 - 图切分：负载均衡
 - 计算容错：自动回复
 - 基于图模型的深度学习算法

知识图谱

- 天猫精灵：语音交互
- 商品推荐与“凑单”
- 挑战
 - 知识抽取和图谱构建：人工较多
 - 高并发/低延时图查询

相关工作

- 图数据库
 - Neo4j
 - Apache Titan
- 批处理系统
 - Google Pregel
 - Apache Giraph
 - Spark GraphX
 - Tux2
 - PowerGraph
- 实时/流计算系统
 - Apache Flink

外来展望

- 根据真实场景，打造测试标准
- 构建一站式图计算平台
- 增强编程语言的表达能力
- 设计合理的分布式软件架构和系统抽象

总结

1. 这篇更偏向于应用层，许多pattern等很有实际意义，实际需求很重要。
2. 更自定义，针对性更强，针对数据，场景，需求定义算法。
3. 开始和人打交道，可视化的作用就凸显了出来，很多研究要和实际的需求结合起来，两个模式匹配的pattern就是有实际意义，可以讲故事，可以针对性的设计算法，其中的参数都有具体的意义，但是衡量起来就很困难，这个地方还是很需要可视化的辅助。

大图数据分析系统综述

常见分析问题

- 搜索
 - bfs
 - dfs
 - 最短路径
- 社区
 - K-core
 - K-truss
 - K-clique
- 节点重要性和相关性
 - PageRank
 - SimRank
 - RandomWalk
- 匹配分析查询

低层次抽象的提供编程接口的图数据分析系统

- Pregel
 - Giraph
 - PowerGraph
 - GraphLab
 - Quegel
 - PAGE
- Trinity
- GraphX

高层次抽象的提供描述性语言的图数据分析系统

- Neo4j
- EmptyHeaded
- gStore

基于图的RDF知识图谱数据管理和分析方法

- RDF
 - gStore
 - gStoreD
 - gAnswer

总结

1. 这篇只是一些简单的介绍，深度的不是很多。
2. 知识图谱讲的更多一点，其中有一些nlp的一些东西